УДК 519.237

Л. Б. КАЩЕЕВ, канд. техн. наук,
Т. Н. ШКАФЕНКО, студентка НТУ «ХПИ»

КЛАСТЕРИЗАЦИЯ ЭКОЛОГИЧЕСКОЙ ЗАГРЯЗНЕННОСТИ ПРОМЫШЛЕННЫХ И СОЦИАЛЬНЫХ ОБЪЕКТОВ

В статті розглянута задача розподілу на класи екологічно небезпечних об'єктів промислової зони з використанням алгоритмів дендритів та навчання без вчителя. Розраховуються функціонали якості розподілу та робляться висновки про їх переваги та недоліки для цієї задачі.

В статье рассмотрена задача кластеризации экологически опасных объектов промышленной зоны с использованием алгоритмов дендритов и обучения без учителя. Рассчитаны функционалы качества разбиения и сделаны выводы про их преимущества и недостатки для данной задачи.

In the article the task of breaking up is considered on classes ecologically dangerous objects of industrial area with the use of algorithms of dendrites and teaching without a teacher. The functional of quality of breaking up are expected and the proper conclusions are done about their advantages and failings for this task.

Введение. Загрязнение окружающей среды в настоящее время является одной из важнейших проблем. Для ее решения необходимо решить задачу одновременного контроля множества факторов загрязнения, каждый из которых ограничен соответствующей санитарной нормой (СНИП). Тем не менее, контролировать каждый из этих факторов достаточно дорого. Поэтому на основании многолетних наблюдений желательно выделить некоторые районы, области, участки улиц, цеха на предприятиях, в которых экологическая обстановка вызывает опасение. Разбиение по уровню загрязнения, позволяет оценить источники загрязнения и возможные штрафные санкциями, применяемыми к предприятиям.

Автоматизация процесса разбиения на классы предполагает реализацию следующего алгоритма – при каждом вводе новых данных перенос объектов из одного класса в другой. Подобные задачи относятся к разделу прикладной математики, именуемому «кластерный анализ». Классы, полученные после разбиения, могут служить критерием того, как часто необходимо производить проверки объектов.

Постановка задачи. Рассматриваются контрольные точки, на которых производились измерения и источники загрязнения. Количество точек ограничено 2,5–3 тысячами. В каждой точке контролируется не менее двух факторов загрязненности. Требуется разбить точки на n классов близкой загрязненности ($n > 2$).

В качестве входных данных используются реальные данные по бензину, керосину, сероводороду в контрольных точках, а также о влиянии на них источников загрязнения. Исходные данные хранятся в базе данных (dbf-

файл). Пользователь устанавливает количество классов разбиения, если необходимо точки, входящие в обучающую выборку. На выходе получается цифровой атрибут принадлежности к тому или иному классу загрязненности. Группирование объектов по числовым параметрам, физический смысл и приоритеты которых определены санитарными нормами.

Подготовка данных к кластеризации. Числовые данные формируются в результате системы запросов из таблиц базы данных. Признаки, включенные в матрицу наблюдений, неоднородны, поскольку описывают разные свойства объектов. Данные нормализуются, то есть распределяются в промежутке от 0 до 1 по формуле (1).

$$x_{new} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}. \quad (1)$$

После этого переходят к расчету элементов матрицы расстояний с учетом всех элементов матрицы наблюдений. Понятие расстояния позволяет оценить степень сходства между отдельными реализациями и между классами.

Чаще всего используются: d_1 — евклидово расстояние

$$d_1(\vec{X}_i, \vec{X}_j) = \sqrt{\sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|^2}; \quad (2)$$

d_2 — расстояние по Манхэттену или метрика «городских кварталов» (для этой меры влияние отдельных больших разностей уменьшается, так как они не возводятся в квадрат). Это расстояние вычисляется по формуле:

$$d_2(\vec{X}_i, \vec{X}_j) = \sum_{k=1}^{n} \left| x_{ik} - x_{jk} \right|; \quad (3)$$

d_3 — расстояние Чебышева, это расстояние может оказаться полезным, когда желают определить два объекта как "различные", если они различаются по какой-либо одной координате; оно вычисляется по формуле:

$$d_3(\vec{X}_i, \vec{X}_j) = \max_k \left| x_{ik} - x_{jk} \right|. \quad (4)$$

Описание и практическая реализация алгоритма обучения. В задачу обучения входит постепенное усовершенствование алгоритма разделения предъявляемых объектов на классы. С этой целью отбирают часть предъявляемых объектов и используют их, в процессе обучения для «тренировки» системы.

Массив исходных данных в обучаемой системе состоит из двух частей: обучающей выборки и тестовой выборки, используемой в процессе испытаний. Главная особенность контролируемого метода классификации заключается в непременном наличии «априорных» сведений о

принадлежности к определенному классу каждого вектора измерений, входящего в обучающую выборку. Таким образом, получают множество векторов измерений от источников, принадлежность которых к определенному классу заранее известна. Данный метод является наиболее приемлемым для решения задачи классификации экологически опасных объектов по степени опасности.

Метод дендритов заключается в том, что из дендрита, построенного на единицах разбиваемого множества, удаляется n–1 самых длинных связей. Тем самым получается разбиение дендрита (и единиц множества) на n заранее заданных частей. Полученное разбиение есть в некотором смысле наилучшее, потому что дендриты, построенные на элементах выделенных n частей, характеризуются минимальной суммой образующих их отрезков. Полученные подмножества, следовательно, включают элементы с близкими значениями признаков. В качестве метрики используется Евклидово пространство. В данной работе используется возведенное в квадрат евклидово расстояние. Такое расстояние придает большие веса более отдаленным друг от друга объектам.

Находится остевое дерево минимальной длины. Определяется число вершин, составляется матрица расстояний. Затем выбирается ребро минимальной длины, которое еще не было выбрано, при условии, что оно не образует цикла с уже выбранными. Для этого до построения дерева каждая вершина ***i*** окрашивается в отличный от других цвет. При выборе очередного ребра, где i и j имеют разные цвета, вершина j и все, окрашенные в ее цвет перекрашиваются в цвет i. Таким образом, выбор вершин разного цвета обеспечивает отсутствие циклов. После выбора $n–1$ ребер все вершины получают один цвет. Последняя связь приводит к перекрашиванию всех ребер в один цвет. Длинна этого ребра наибольшая из длин ребер, которые составляют минимальный каркас, поэтому при разбиении на m классов не закрашиваются последние m ребер. В результате вершины, соединенные ребрами различных цветов, принадлежат различным классам.

Для оценки полученного решения использован функционал качества разбиения $F(S)$, определяемый на множестве всех возможных разбиений. Наилучшим разбиением S понимается то, на котором достигается экстремум функционала. В качестве функционалов часто используются такие характеристики:

– средние внутриклассовые расстояния

$$\mathrm{F1(S)} = \left(\sum_{i,\, j \in s_l} d_{ij} \right) / \left(\sum_{l=1}^{k} n_l^2 \right), \qquad (5)$$

– средние межклассовые расстояния.

$$F2(S) = \left(\sum_{i \in s_q, j \in s_q} d_{ij} \right) / \left(\sum_{l<q} n_l \cdot n_q \right). \quad (6)$$

В этих формулах n_l, n_q – число объектов в классах; l, q – номера классов; k – число классов; dij – расстояние между i и j объектами; s_l, s_q – классы.

Таблица данных

Кол-во классов	Минимум функционала F1		Максимум функционала F2	
	Дендриты	Обучение	Дендриты	Обучение
5	0.03	0.02	0.73	0.70
4	0.02	0.01	0.84	0.76
3	0.03	0.03	0.83	0.72

В зависимости от того, чего хочет достичь пользователь (наибольшей компактности – F1, либо наибольшего различия между классами – F2) выбирается соответствующий метод. Для достижения наибольшей компактности лучше использовать метод обучения, для наибольшего различия между классами – метод дендритов.

Выводы. В статье рассмотрена методика группирования многопараметрических объектов. Описана реализация таких методов, как обучение и метод дендритов (Вроцлавская таксономия). Для сравнения методов использованы такие функционалы качества как средние внутриклассовые расстояния и средние межклассовые расстояния. Решена проблема выбора метода для данной экологической задачи.

Список литературы: ***1.*** *Айвазян С. А., Бежаева З. И., Староверов О. В.* Классификация многомерных наблюдений. – М.: Статистика, 1974. – 238 с. ***2.*** *Андерсон Т.* Введение в многомерный статистический анализ. – М.: Физматгиз, 1963. – 500 с. ***3.*** *Кендалл М. Дж., Стьюарт А.* Статистические выводы и связи.– М.: Наука, 1973. – 899 с. ***4.*** *Орлов А. И.* Прикладной многомерный статистический анализ.– М.: Наука, 1973. – 510 с. ***5.*** *Плюта В.* Сравнительный многомерный анализ в экономических исследованиях. – М.: Статистика, 1980. – 150 с.

Поступила в редколлегию 05.04.07